Appendix D. Description of SIPP 1985 Panel File and Data Quality

DESCRIPTION OF SIPP 1985 PANEL FILE

The estimates presented in this report are based on the second SIPP panel file. This file contains monthly data for persons over a 32-month period (28-month period for the first rotation group). The staggered SIPP design (described in appendix A) means that the actual reference periods are January 1985 to April 1987, October 1984 to May 1987, November 1984 to June 1987, and December 1984 to July 1987. The period covered by the 1985 longitudinal panel file consists of 32 interview months (eight interviews) for rotations 2, 3, and 4. Rotation 1 has only 28 interview months (seven interviews). Data from all four rotation groups are available only for the reference period January 1985 through April 1987.

Each person in the panel file has been assigned three weights: a weight for calendar year 1985, a weight for calendar year 1986, and a weight for the 28- or 32-month reference period. In order to receive a nonzero weight, a person must have an observation for each month of the relevant reference period (in this report, 1985 and 1986) or have a complete set of observations up until the time he or she died or became institutionalized. The data shown in this report are affected if characteristics of persons with an incomplete set of observations differed from those with a complete set.

Table D-1 shows three categories of sample persons by sex, age and program participation status. The numbers in the table are unit counts; they are not weighted. The category "complete set of interviews obtained" includes 23,093 persons, but 651 of these persons died or were institutionalized during the 32-month reference period. The next category, "Interviewed in first wave, left sample for reasons other than death or institutionalization" includes 13,620 persons. The final category includes 6,277 persons who were not

Table D-1. **Percent Distribution: Three Categories of Sample Persons**

Characteristic	Complete set of interviews obtained[1]	Interviewed in first wave, left sample for reasons other than death or institutionalization	Not a member of sample household during first wave, interview obtained in second or later waves
Total	23,093	13,620	6,277
	(100.0)	(100.0)	(100.0)
Sex:			
Male	47.5	48.4	50.4
Female	52.5	51.6	49.6
Age at first interview:			
Under 18 years	28.4	27.2	35.0
Under 6 years	10.1	9.1	21.7
18 to 24 years	10.2	14.2	24.4
25 to 44 years	29.6	30.3	26.5
45 to 64 years	19.6	19.0	11.1
65 years and over	12.2	9.4	3.1
75 years and over	4.6	3.6	1.1
Program participation, first month in sample:			
Persons 18 years and over	16,530	9,909	4,082
	(100.0)	(100.0)	(100.0)
Participated in major assistance program	8.8	9.0	9.7
AFDC or general assistance	2.1	2.7	3.1
Food stamps	4.7	4.8	4.7
Medicaid	5.0	4.8	5.3
Public/subsidized housing	3.0	3.1	2.4
SSI	2.2	1.5	1.6
Did not participate	91.2	91.0	90.3

[1]Includes 651 persons who died or were institutionalized during the 32-month period.

a member of a SIPP household during the first wave of interviews, but who subsequently became a member of a sample household.

A comparison of the first two columns shows the characteristics of those who completed the full set of interviews are reasonably close to the characteristics of those who dropped out of the sample. The major differences in the age distribution are for young adults and for the elderly. Young adults are underrepresented and the elderly are overrepresented in the group of persons who completed the full set of interviews. The data in table D-1 are, as noted, unweighted, and any potential problem caused by unrepresentative age distributions are minimized when the file is weighted to independent controls.

TIME-IN-SAMPLE BIAS

The use of the panel file to obtain estimates for 1985 and 1986 raises the issue of time-in-sample bias. There is ample evidence that certain measures vary according to the number of times the respondent has been visited. In the CPS, for example, the measured unemployment rate is always higher for the group of households being interviewed for the first time than for the groups being interviewed for the second or later times.

Time-in-sample bias arises when a person's response to a survey question (or the interviewer's method of asking a question) is influenced by what occurred in a previous visit. The overlapping SIPP sample design provides the data that allows for an examination of the presence of time-in-sample bias in SIPP estimates. That is, it is possible in SIPP to obtain estimates for a given time period from two or more separate panels and the amount of time respondents will have spent in the SIPP panel will differ for each of the panels. For example, estimates for each of the four quarters of 1986 can be obtained from both the 1984 and 1985 panels (respondents in the 1984 will have had more visits).

The quarterly estimates in table D-2 are shown for the four quarters of 1985 and for the first quarter of 1986. Estimates from each panel file are shown separately for comparison. The estimates shown are of median income of nonfarm households, number of households receiving Social Security or Railroad Retirement, number of households receiving food stamps, and number of households with low monthly income.

The figures in table D-2 provide very little evidence regarding the existence of time-in-sample bias for several reasons. First, most of the observed differences are smaller than the differences that could be explained by sampling error. Second, a single observation is not sufficient to identify a pattern of bias. Third, differences may be attributable to attrition bias rather than time-in-sample bias. In spite of these qualifications, however, the observed relationships offer some reason to be cautious in interpreting the differences that have been presented earlier in this report—both the differences between CPS and SIPP estimates and the differences between the 1984 and 1985 estimates that were obtained from the SIPP.

OTHER ISSUES OF DATA QUALITY

Two major determinants of the quality of income data collected in household surveys are the magnitude of missing responses and the accuracy of the responses that are provided. This appendix has been included to supply information concerning nonresponse rates for selected income questions, the average amounts of income reported in the survey or assigned in the imputation of missing responses, and the extent to which the survey figures underestimate numbers of income recipients and amounts of income received.

Nonresponse in this discussion refers to missing responses to specific questions or "items" on the questionnaire. Noninterviews or complete failure to obtain cooperation from any household member have not been considered in this examination of nonresponse rates. Adjustments to account for noninterviews are made by proportionally increasing the survey weights of interviewed households. Missing responses to specific questions are assigned a value in the imputation phase of the data processing operation.

Nonresponse is a very important factor in assessing the quality of survey data. Nonresponses to income questions cannot be considered random since experience has shown that persons with the highest nonresponse rates have reported characteristics such as education levels and occupations that, in general, differ from population averages. The most frequent causes of nonresponse are the inability of the respondent to answer the question because of either a 1) lack of knowledge or 2) refusal to answer. The first reason is especially important in situations of proxy response when one household member answers questions for another household member not present at the time of the interview. The practice of accepting proxy interviews from household members deemed "qualified" to answer is a standard procedure in the CPS and most other surveys conducted by the Bureau. During the eight interviews of the SIPP 1985 panel, an average of 36 percent of the interviews were taken from proxy respondents.

Nonresponses are assigned values prior to producing estimates from the survey data. The procedure used to assign or impute responses for missing data for SIPP are of a type commonly referred to as a "hot deck" imputation method. This process assigns values reported in the survey by respondents to nonrespondents. The respondent from whom the value is taken is termed the "donor." Values from donors are stored in a matrix

Table D-2. **Selected Monthly Averages, by Quarter: 1984 and 1985 SIPP Cross-Sectional Files**

Characteristic	Source of estimate				1984 panel to 1985 panel
	1984 panel	Standard error	1985 panel	Standard error	
Median income of nonfarm households:					
1985, quarter 1	$1,811	$20	$1,790	$21	1.01
1985, quarter 2	1,861	21	1,838	22	1.02
1985, quarter 3	1,858	22	1,855	23	1.00
1985, quarter 4	1,891	22	1,886	24	1.00
1986, quarter 1	1,887	22	1,897	24	0.99
Number of households receiving Social Security or Railroad Retirement (thous.):					
1985, quarter 1	23,821	385	23,559	403	1.01
1985, quarter 2	23,955	386	23,781	405	1.01
1985, quarter 3	23,938	386	23,838	405	1.00
1985, quarter 4	23,864	385	23,929	405	1.00
1986, quarter 1	23,867	385	24,145	407	0.99
Number of households receiving food stamps (thous.):					
1985, quarter 1	6,230	223	5,999	229	1.04
1985, quarter 2	5,955	218	5,808	226	1.03
1985, quarter 3	5,886	217	5,624	223	1.05
1985, quarter 4	5,839	216	5,676	224	1.03
1986, quarter 1	5,965	218	5,800	226	1.03
Number of households with low monthly income (thous.):					
1985, quarter 1	10,922	286	11,585	308	0.94
1985, quarter 2	10,783	285	10,929	300	0.99
1985, quarter 3	10,872	286	11,088	302	0.98
1985, quarter 4	10,688	284	10,978	301	0.97
1986, quarter 1	10,878	286	10,890	300	1.00

defined by demographic and economic data available for both donors and nonrespondents. Each cell of the matrix defines a unique combination of demographic and economic characteristics. For example, the imputation of an amount for monthly wage and salary income is based on eight different variables. These were 1) occupation, 2) sex, 3) age, 4) race, 5) educational attainment, 6) weeks worked, 7) usual hours worked per week, and 8) place of residence.

The second important determinant of data quality and probably the one examined most closely by users of the income data collected in household surveys is the accuracy of reported (and imputed) amounts. In general, household surveys have a tendency to underestimate the number of persons receiving income and the average amount received. These problems result for a variety of reasons including random response error, misreporting of sources of income, failure to report the receipt of income from a specified source, and failure to report the full amount received. The net effect of these kinds of problems is, for most income types, underestimation or underreporting of income amounts. The extent of underreporting is measured by comparing survey estimates with independently derived estimates, usually based on administrative data that are, generally, more reliable than the estimates derived from the survey. It should be noted that the independent estimates are subject to errors themselves. In addition, independent estimates do not reflect income attributable to the "underground" economy, some of which may be reported in the survey.